Minimax Prediction for Functional Linear Regression with Functional Responses in Reproducing Kernel Hilbert Spaces

Heng Lian

Division of Mathematical Sciences, School of Physical and Mathematical Sciences,

Nanyang Technological University, Singapore, 637371

Abstract

In this article, we consider convergence rates in functional linear regression with functional responses, where the linear coefficient lies in a reproducing kernel Hilbert space (RKHS). Without assuming that the reproducing kernel and the covariate covariance kernel are aligned, or assuming polynomial rate of decay of the eigenvalues of the covariance kernel, convergence rates in prediction risk are established. The corresponding lower bound in rates is derived by reducing to the scalar response case. Simulation studies and two benchmark datasets are used to illustrate that the proposed approach can significantly outperform the functional PCA approach in prediction.

keywords: Functional data; Functional response; Minimax convergence rate; Regularization.

1 Introduction

The literature contains an impressive range of functional analysis tools for various problems including exploratory functional principal component analysis, canonical correlation analysis, classification and regression. Two major approaches exist. The more traditional approach, masterfully documented in the monograph (Ramsay and Silverman, 2005), typically starts by representing functional data by an expansion with respect to

a certain basis, and subsequent inferences are carried out on the coefficients. The most commonly utilized basis include B-spline basis for nonperiodic data and Fourier basis for periodic data. Another line of work by the French school (Ferraty and Vieu, 2002), taking a nonparametric point of view, extends the traditional nonparametric techniques, most notably the kernel estimate, to the functional case. Some recent advances in the area of functional regression include Cardot et al. (2003); Cai and Hall (2006); Preda (2007); Lian (2007); Ait-Saidi et al. (2008); Yao et al. (2005); Crambes et al. (2009); Ferraty et al. (2011); Lian (2011).

In this paper we study the functional linear regression problem of the form

$$Y(t) = \mu(t) + \int_0^1 \beta(t,s)X(s)\,ds + \epsilon(t), \tag{1}$$

where $Y, X, \epsilon \in L_2[0,1]$ and $E[\epsilon|X] = 0$, the same problem that appeared in Ramsay and Silverman (2005); Yao et al. (2005); Antoch et al. (2008); Aguilera et al. (2008); Crambes and Mas (2012). In terms of methodology, the plan of attack we will give for (1) is most closely related to that of Crambes and Mas (2012). In this introduction, we will explain the methodology used in that paper and then the different assumption we will make on $\beta(t,s)$.

Without loss of much generality, throughout the paper we assume $E(X) = 0$ and the intercept $\mu(t) = 0$, since the intercept can be easily estimated. The covariance operator of X is the linear operator $\Gamma = E(X \otimes X)$ where for $x, y \in L_2[0,1]$, $x \otimes y : L_2[0,1] \to L_2[0,1]$ is defined by $(x \otimes y)(g) = \langle y, g\rangle x$ for any $g \in L_2[0,1]$. Γ can also be represented by the bivariate function $\Gamma(s,t) = E[X(s)X(t)]$. Using the same letter Γ to denote both the operator and the bivariate function will not cause confusion in our context. We assume throughout the paper that $E\|X\|^4 < \infty$ which implies Γ is a compact operator. Then by the Karhunen-Loève Theorem there exists a spectral expansion for Γ,

$$\Gamma = \sum_{j=1}^{\infty} \lambda_j \varphi_j \otimes \varphi_j,$$

where $\lambda_j \geq 0$ are the eigenvalues with $\lambda_j \to 0$ and $\{\varphi_j\}$ are the orthonormalized eigenfunctions. Correspondingly, we have the representation $X = \sum_{j\geq 1} \gamma_j \varphi_j$ with $\gamma_j = \int X\varphi_j$. The random coefficients γ_j satisfies $E\gamma_j\gamma_k = \lambda_j I\{j = k\}$ where $I\{.\}$ is the indicator function.

By expanding β using the set of eigenfunctions, we write $\beta(t,s) = \sum_{j\geq 1} b_j(t)\varphi_j(s)$ and (1) can be equivalently written as

$$Y(t) = \sum_{j\geq 1} b_j(t)\gamma_j + \epsilon(t).$$

Multiplying both sides above by γ_j and taking expectations, we easily obtain $b_j(t) = E[Y(t)\gamma_j]/\lambda_j$. Given i.i.d. data $(X_i, Y_i), i = 1, \ldots, n$, $\{\lambda_j, \varphi_j\}$ can be easily estimated by $\hat{\lambda}_j$ and $\hat{\varphi}_j$ obtained from the spectral decomposition of the empirical covariance operator and $E[Y(t)\gamma_j]$ can be approximated by the corresponding sample average. Thus the estimator proposed in Crambes and Mas (2012) is

$$\hat{\beta}(t,s) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{\int X_i\hat{\varphi}_j}{\hat{\lambda}_j}Y_i(t)\hat{\varphi}_j(s).$$

Note that the infinite sum over j has been truncated as some point k for regularization. One intriguing point is that there is no regularization on $Y_i(t)$ necessary, in contrast with Yao et al. (2005) where Y is observed sparsely with additional noise. This can also be seen from that $b_j(t)$ is not a priori constrained in any way. The reason is that only regularization of the covariance operator, which does not depend on Y, is necessary to avoid overfitting.

Minimax convergence rates of $E\|\int \hat{\beta}(t,s)X(s)ds - \int \beta(t,s)X(s)ds\|^2$ were shown in Crambes and Mas (2012). A key assumption is the appropriate decaying assumption on $\|b_j\|$ as j increases. Given that $\|b_j\|$'s are the coefficients of $\beta(t,s)$ in terms of the basis φ_j, which is a characteristic of the predictor, there is no a priori reason why this basis should provide a good representation of β in the sense that $\|b_j\|$ will decay fast. Indeed, a more reasonable assumption for β is on its smoothness, which makes a reproducing kernel Hilbert space (RKHS) approach more reasonable conceptually. Such arguments have led to the developments in Yuan and Cai (2010); Cai and Yuan (2012) for the scalar response models. While Crambes and Mas (2012) is based on Cardot et al. (2007) for scalar response models, ours is based on Cai and Yuan (2012).

The rest of the article is organized as follows. In Section 2, we propose an estimator for β with an RKHS approach where the reproducing kernel and the covariance kernel are not necessarily aligned. We establish the minimax rate of convergence in

prediction risk by deriving both the upper bound and the lower bound. In Section 3, we present some simulation studies to show that the RKHS approach could significantly outperform the functional PCA approach when the kernels are mis-aligned. This advantage is further illustrated on two benchmark datasets which shows better prediction performance using our approach. We conclude in Section 4 with some discussions. The technical proofs are relegated to the Appendix.

Finally, we list some notations and properties regarding different norms to be used. For any operator $\mathcal{F}$, we use $\mathcal{F}^{\mathrm{T}}$ to denote its adjoint operator. If $\mathcal{F}$ is self-adjoint and nonnegative definite, $\mathcal{F}^{1/2}$ is its square-root satisfying $\mathcal{F}^{1/2}\mathcal{F}^{1/2} = \mathcal{F}$. For $f \in L_2$, $\|f\|$ denotes its L_2 norm. For any operator $\mathcal{F}$, $\|\mathcal{F}\|_{op}$ is the operator norm $\|\mathcal{F}\|_{op} := \sup_{\|f\|\le 1} \|\mathcal{F}f\|$. The trace norm of an operator $\mathcal{F}$ is $\mathrm{Trace}(\mathcal{F}) = \sum_k \langle (\mathcal{F}^{\mathrm{T}}\mathcal{F})^{1/2} e_k, e_k\rangle$ for any orthonormal basis $\{e_k\}$ of L_2. $\mathcal{F}$ is a trace class operator if its trace norm is finite. The Hilbert-Schmidt norm of an operator is $\|\mathcal{F}\|_{HS} = (\sum_{j,k}\langle \mathcal{F}e_j, e_k\rangle^2)^{1/2} = (\sum_j \|\mathcal{F}e_j\|^2)^{1/2}$. An operator is a Hilbert-Schmidt operator if its Hilbert-Schmidt norm is finite. From the definition it is easy to see that $\mathrm{Trace}(\mathcal{F}^{\mathrm{T}}\mathcal{F}) = \mathrm{Trace}(\mathcal{F}\mathcal{F}^{\mathrm{T}}) = \|\mathcal{F}\|_{HS}^2$, Furthermore, if $\mathcal{F}$ is a Hilbert-Schmidt operator and $\mathcal{G}$ is a bounded operator, then $\mathcal{F}\mathcal{G}$ is also a Hilbert-Schmidt operator with $\|\mathcal{F}\mathcal{G}\|_{HS} \le \|\mathcal{F}\|_{HS}\|\mathcal{G}\|_{op}$.

2 Methodology and Convergence Rates

Following Wahba (1990), a RKHS H is a Hilbert space of real-valued functions defined on, say, the interval $[0,1]$, in which the point evaluation operator $L_t : H \to R, L_t(f) = f(t)$ is continuous. By Riesz representation theorem, this definition implies the existence of a bivariate function $K(s,t)$ such that

$$\begin{aligned}
&K(s,\cdot) \in H, \text{ for all } s \in [0,1]\\
&\text{and (reproducing property)}\\
&\text{for every } f \in H \text{ and } t \in [0,1],\ \langle K(t,\cdot), f\rangle_H = f(t).
\end{aligned}$$

The definition of a RKHS can actually start from a positive definite bivariate function $K(s,t)$ and RKHS is constructed as the completion of the linear span of $\{K(s,\cdot), s \in [0,1]\}$ with inner product defined by $\langle K(s,\cdot), K(t,\cdot)\rangle_H = K(s,t)$. To make the dependence on K explicit, the RKHS is denoted by H_K with the RKHS norm $\|\cdot\|_{H_K}$. With

abuse of notation, K also denotes the linear operator $f \in L_2 \to Kf = \int K(\cdot, s)f(s)ds$. For later use, we note that H_K is identical to the range of $K^{1/2}$.

We assume that for any $t \in [0,1]$, $\beta(t,\cdot) \in H_K$. This is a smoothness assumption for $\beta(t,s)$ in the s-variable. As noted in the introduction, smoothness assumption on the t-variable is not necessary. We estimate β via

$$\hat{\beta} = \arg \min_{\beta(t,.) \in H_K} \frac{1}{n} \sum_{i=1}^{n} \|Y_i - \int_0^1 \beta(\cdot, s) X_i(s)\, ds\|^2 + \lambda \int_0^1 \|\beta(t,\cdot)\|_{H_K}^2 dt. \qquad (2)$$

We implicitly assume that the expression $\int_0^1 \|\beta(t,\cdot)\|_{H_K}^2 dt$ is valid, that is $\|\beta(t,\cdot)\|_{H_K}$ as a function of t is square integrable. This assumption on β is also more succinctly denoted by $\beta \in L^2 \times H_K$.

The following representer theorem is useful in computing the solution, whose proof is omitted since it is standard.

Proposition 1 *The solution of (2) can be expressed as*

$$\hat{\beta}(t,s) = \sum_{i=1}^{n} c_i(t) \int_0^1 K(s,u) X_i(u)\, du. \qquad (3)$$

Based on the previous proposition, by plugging the representation (3) into (2), it can be easily shown that $(c_1(t), \ldots, c_n(t))^{\mathrm{T}} = (\Sigma + n\lambda)^{-1} Y(t)$ where Σ is an $n \times n$ matrix whose entries are given by $\Sigma_{ij} = \int \int X_i(s) K(s,t) X_j(t) ds dt$.

Remark 1 *Throughout this section, we assume the reproducing kernel K is positive definite and the RKHS norm for H_K is used in the penalty. More generally, for practical use, we can assume $H_K = H_1 \oplus H_2$, where H_1, typically finite dimensional, is a RKHS with reproducing kernel K_1 and H_2 is a RKHS with reproducing kernel K_2, $K = K_1 + K_2$. We can then impose the penalty $\int \|P_2 \beta(t,\cdot)\|_{H_K}^2 dt = \int \|P_2 \beta(t,\cdot)\|_{H_2}^2 dt$, where P_2 is the projection onto H_2. Our theory and computation can be easily adapted to this more general case, but we use (2) for ease for presentation throughout the paper. In real data analysis, $H_K = \mathcal{W}_2^{per}$ is the second-order Sobolev space of periodic functions on $[0,1]$ and we use decomposition $H_K = H_1 \oplus H_2$ where H_1 contains the constant functions.*

Since $\beta(t,\cdot) \in H_K$, there exists $f(t,s)$ such that $\beta(t,\cdot) = K^{1/2} f(t,\cdot)$ and $\|\beta(t,\cdot)\|_{H_K} = \|f(t,\cdot)\|$. Thus (2) can also be written as

$$\hat{f} = \arg \min_{f \in L_2[0,1]^2} \frac{1}{n} \sum_{i=1}^{n} \|Y_i - \int_0^1 f(\cdot, s)(K^{1/2} X_i)(s)\, ds\|^2 + \lambda \int_0^1 \int_0^1 f^2(t,s) ds dt. \quad (4)$$

Due to the appearance of $K^{1/2} X_i$ in the expression above, this suggests that the spectral decomposition of $T := K^{1/2} \Gamma K^{1/2}$ plays an important role. Suppose the spectral decomposition of T is

$$T = \sum_{j \geq 1} s_j e_j \otimes e_j,$$

with $s_1 > s_2 > \cdots > 0$.

The following technical assumptions are imposed.

(A1) There exists a positive, convex, decreasing function $\phi : (0, \infty) \to R^+$ such that $s_j = \phi(j)$ at least for large j.

(A2) Recall the Karhunen-Loéve expansion $K^{1/2} X = \sum_{j \geq 1} \xi_j e_j$. There exists a constant c such that $E[\xi_j^4] \leq c(E[\xi_j^2])^2$ for all $j \geq 1$.

(A3) $\beta(t,\cdot) \in H_K$ for all $t \in [0,1]$, and $\|\beta(t,\cdot)\|_{H_K} \in L_2$ as a function of t. Furthermore, $BK^{-1/2}$ is a Hilbert-Schmidt operator, where the operator B is defined by $Bf = \int \beta(\cdot, s) f(s) ds, f \in L_2$.

Assumption (A1) also appeared in Cardot et al. (2007). Cai and Yuan (2012) considered a much more restrictive polynomial decay assumption $s_j \asymp j^{-2r}$ for some $r > 0$, which corresponds to $\phi(x) = x^{-2r}$. Taking $\phi(x) = c_1 e^{-c_2 x}$ for some constants $c_1, c_2 > 0$, exponential decay of eigenvalues is also a special case of our result, among many others.

Assumption (A2) is similar to that assumed in Hall and Horowitz (2007); Cardot et al. (2007). Cai and Yuan (2012) assumed that $E(\int X(t) f(t) dt)^4 \leq c(E(\int X(t) f(t) dt)^2)^2$ for all $f \in L_2$. This assumption implies (A2) which can be seen by choosing $f = K^{1/2} e_j$.

(A3) is a natural extension of the case with scalar reponse, where $\beta(t) \in H_K$ automatically implies $K^{-1/2} \beta \in L_2$. Superficially, $BK^{-1/2}$ in (A3) is only defined on the

range of $K^{1/2}$, which coincides with H_K and is a dense subset of L_2. Also, since $K^{-1/2}$ is an unbounded operator, it is not clear that $BK^{-1/2}$ can be bounded. Nevertheless, it can be shown that under the condition that $\beta(t,\cdot) \in H_K$ and $\|\beta(t,\cdot)\|_{H_K} \in L_2$, $BK^{-1/2}$ is bounded on L_2. More specifically, we have the following proposition whose proof is in the Appendix.

Proposition 2 *If $\beta(t,\cdot) \in H_K$ for all $t \in [0,1]$ and $\|\beta(t,\cdot)\|_{H_K} \in L_2$ where $\|\beta(t,\cdot)\|_{H_K}$ is regarded as a function of t, then $BK^{-1/2}$ is a bounded operator on L_2.*

The risk we consider is the prediction risk $E^*\|\hat{B}(X^*) - B(X^*)\|^2$ where X^* is a copy of X independent of the training data and E^* is the expectation taken over X^*. We first present the upper bound.

Theorem 1 *Under assumptions (A1)-(A3), and that $\lambda \to 0, \lambda n \to \infty$, we have*

$$E^*\|\hat{B}(X^*) - B(X^*)\|^2 = O_p\left(\lambda + \frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2}\right).$$

Remark 2 *By examining the proof carefully, one can actually see that the convergence is uniform in β that satisfies (A3) with $\|BK^{-1/2}\|_{HS} \le 1$ (there is nothing special about the upper bound 1, which can be replace by any $L > 0$). We can thus actually show*

$$\lim_{a\to\infty}\lim_{n\to\infty}\sup_{\beta\in L_2\times H_K, \|BK^{-1/2}\|_{HS}\le 1} P(E^*\|\hat{B}(X^*) - B(X^*)\|^2 > a\lambda_0) = 0$$

This expression is put here for easy comparison with the lower bound obtained in Theorem 2 below.

We now discuss how to choose appropriate λ to balance the two terms in the rate above. Let $J = \lfloor \phi^{-1}(\lambda) \rfloor$ be the integer part of $\phi^{-1}(\lambda)$. By splitting the sum over j into $j \le J$ and $j > J$, we have

$$\frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2} \le \frac{J}{n} + \frac{s_{J+1}\sum_{j\ge J+1} s_j}{n\lambda^2}.$$

Let λ_0 be the solution to the equation

$$\phi^{-1}(\lambda) = n\lambda. \tag{5}$$

Then we have $J_0 := \lfloor \phi^{-1}(\lambda_0) \rfloor \leq \phi^{-1}(\lambda_0)$ and

$$\frac{s_{J_0+1} \sum_{j \geq J_0+1} s_j}{n\lambda_0^2} \leq \frac{(J_0+2)s_{J_0+1}^2}{n\lambda_0^2} \leq \frac{J_0+2}{n},$$

where we used that $\sum_{j \geq J_0+1} s_j \leq (J_0+2)s_{J_0+1}$ obtained from Lemma 1 of Cardot et al. (2007), and that $s_{J_0+2} = \phi(J_0+2) \leq \phi(\phi^{-1}(\lambda_0)) = \lambda_0$ by the definition of J_0. Thus we have

$$E^* \|\hat{B}(X^*) - \hat{B}(X^*)\|^2 = O_p(\lambda_0)$$

with λ_0 defined by (5), which characterizes the optimal convergence rate. In the special case $\phi(x) = x^{-2r}$, $\lambda_0 = n^{-2r/(2r+1)}$, which is the same as the rate obtained in Cai and Yuan (2012) for scalar response models. On the other hand, if $\phi(x) = e^{-x}$, we can easily show that $\text{loglog}n/n < \lambda_0 < \text{log}n/n$, an almost parametric rate.

We now establish the lower bound. This is obtained by first reducing the problem to the scalar response model and then using a slightly different construction from that used in Cai and Yuan (2012) to deal with more general ϕ. The details of the proof are contained in the Appendix.

Theorem 2 *Under assumptions (A1) and (A2) on the predictor distribution, we have, for any $a > 0$*

$$\lim_{a \to 0} \lim_{n \to \infty} \inf_{\hat{\beta}} \sup_{\beta \in L_2 \times H_K, \|BK^{-1/2}\|_{HS} \leq 1} P(E^* \|\hat{B}(X^*) - B(X^*)\|^2 > a\lambda_0) = 1$$

where the infimum is taken over all possible estimators based on the training data $(X_i, Y_i), i = 1, \ldots, n$.

3 Numerical Results

3.1 Simulations

The simulation setup is similar to that used in Cai and Yuan (2012). We consider the RKHS with kernel

$$K(s,t) = \sum_{j\geq 1} \frac{2}{(j\pi)^4} \cos(j\pi s)\cos(j\pi t),$$

and thus H_K consists of functions of the form

$$f(t) = \sum_{j\geq 1} f_j \cos(j\pi t)$$

such that $\sum_j j^4 f_j^2 < \infty$. In this case, we actually have $\|f\|_{H_K}^2 = \int (f'')^2$. Data are generated from (1) without the intercept term, with

$$\beta(t,s) = \sum_j 4\sqrt{2}(-1)^j \frac{\sin(j\pi t)}{j^2} \cos(j\pi s).$$

For the covariance kernel, we use

$$\Gamma(s,t) = \sum_{j\geq 1} 2\theta_j \cos(j\pi s)\cos(j\pi t),$$

where $\theta_j = (|j - j_0| + 1)^{-2}$. When $j_0 = 1$, the two kernels are perfectly aligned, in the sense that they have the same sequence of eigenfunctions when ordered according to the eigenvalues. As j_0 increases, the level of mis-alignment also increases and we expect that the performance of functional PCA approach deteriorate with j_0. After finding the integral $Z_i(t) := \int \beta(t,s)X_i(s)ds$ (approximated easily by a Riemannian sum), we discretize Z_i over $[0,1]$ on an equally-spaced grid $(t_1,\ldots,t_{100})$ with 100 points and then add independent $\epsilon_{ik} \sim N(0,\sigma^2)$ noises to finally obtain $Y_i(t_k) = Z_i(t_k) + \epsilon_{ik}$. The discretized data for model fitting contains $(X_i(t_k), Y_i(t_k)), k = 1,\ldots,100, i = 1,\ldots,n$. We set $n = 50, 100$ and $\sigma = 0.1, 0.3$, resulting in a total of four scenarios for each j_0. For values of j_0, we use $j_0 \in \{1,3,5,\ldots,15\}$. For the functional PCA approach, the tuning parameter is the truncation point which we

consider in the range from 1 to 25. For the RKHS approach, the tuning parameter is λ and we consider $\lambda \in \exp\{-20, -19, \ldots, 0\}$. The experiment for each scenario was repeated 100 times.

In this simulation, the tuning parameters are chosen to yield the smallest error to reflect the best achievable performance for both methods. To assess the performance, 100 test predictors $X_1^*, \ldots, X_{100}^*$ are generated from the same model as the training data, and root mean squared error (RMSE) is defined to be $(\sum_i \| \int \hat{\beta} X_i^* - \int \beta X_i^* \|^2 / 100)^{-1/2}$. Simulation results are summarized in Figure 1, which shows the RMSE for both methods. Each panel corresponds to a pair of values of (n, σ), and the curves show the RMSE averaged over 100 replications for both methods as j_0 increases (red curve for the functional PCA approach and black curve for the RKHS approach). The vertical bar shows $\pm$ 2 standard errors computed from the 100 replications.

It is clearly seen that the performance of the RKHS approach is similar to (actually better than) that of the functional PCA approach for $j_0 = 1$. As j_0 increases, the performance of the functional PCA approach becomes much worse, while the errors for the RKHS approach remain at the same level. The difference in performance between these two methods generally increases with j_0 (with some exceptions in our particular simulations).

3.2 Real data

We now turn to the prediction performance of the proposed method on two real datasets. These datasets are used frequently in functional data analysis, and both are available from the fda package in R.

Canadian weather data. The daily weather data consists of daily temperature and precipitation measurements recorded in 35 Canadian weather stations. Each observation consists of functional data observed on an equally-spaced grid of 365 points. We treat the temperature as the independent variable and the goal is to predict the corresponding precipitation curve given the temperature measurements. As is previously done, we set the dependent variable to be the log-transformed precipitation measurements, and a small positive number is added to the values with 0 precipitation recorded. Given the periodic nature of the data, we set $H_K = \mathcal{W}_2^{per}$, the second-order Sobolev space of periodic functions on $[0, 1]$. The reproducing kernel is given by

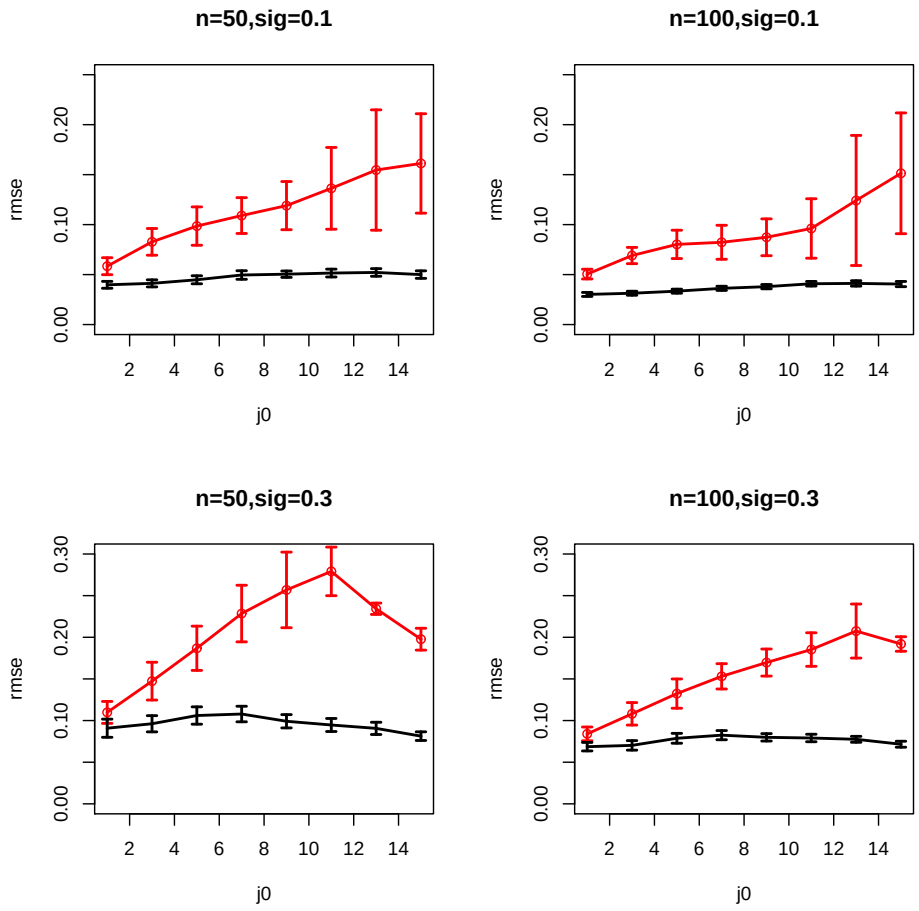

Figure 1: RMSE for both the functional PCA method (red curve) and the RKHS method (black curve) for the simulated data using the optimal tuning parameters.

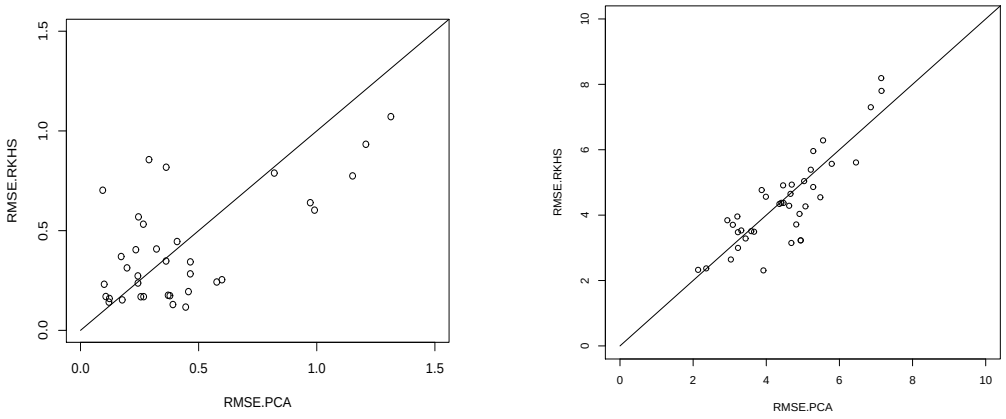

Figure 2: Leave-one-out prediction error for the real data. The x-coordinate for each point shows the error of the functional PCA method, and the y-coordinate shows the error of the RKHS method. Left panel: Canadian weather data; Right panel: Gait data. The tuning parameters are chosen to minimize the leave-one-out cross-validation error in both methods.

$K(s,t) = K_1(s,t) + K_2(s,t)$ with $K_1(s,t) = 1, K_2(s,t) = \sum_{j\geq 1} \frac{2}{(2\pi j)^4} \cos(2\pi j(s-t))$. The modification as mentioned in Remark 1 is used. We use leave-one-out cross-validation to determine the best tuning parameters to use for both methods. The left panel in Figure 2 shows the prediction errors on the 35 stations using the best tuning parameters (trained on 34 stations). For 20 stations, the functional PCA approach has larger error than the RKHS approach. The average mean prediction error for the functional PCA approach is 0.43 while the error is 0.40 for the RKHS approach.

Gait data. The Motion Analysis Laboratory at Children's Hospital, San Diego, collected these data, which consist of the angles formed by the hip and knee of 39 children over each child's gait cycle. The cycle begins and ends at the point where the heel of the limb under observation strikes the ground. Both sets of functions are periodic and it is of interest to see how the two joints interact. In this application, we use hip angle as the predictor and knee angle as the response. The right panel in Figure 2 shows the prediction errors on the 39 children. For 21 children, the functional PCA approach has larger error than the RKHS approach. The average mean prediction error for the functional PCA approach is 4.49 while the error is 4.38 for the RKHS approach.

4 Conclusion

In this paper, we established the minimax rate of convergence for prediction in functional response models in the general setting where the covariance kernel Γ and the reproducing kernel K are not aligned, and also under general assumption on the decay rate of the eigenvalues of operator $T = K^{1/2}\Gamma K^{1/2}$. Our simulations show that as the degree of alignment of the two kernels decreases, the RKHS estimator can significantly outperform the estimator based on functional PCA. The two real datasets further demonstrate that the RKHS estimator can have better prediction accuracy.

Choice of tuning parameter λ can be done via cross-validation, as illustrated in our analysis of the real data. Cai and Yuan (2012) proposed an adaptive method for tuning parameter selection which is an important theoretical development, but in our experience does not work as well as cross-validation. Theoretical development of a good tuning parameter selector can be of significant importance which we do not investigate here.

Furthermore, one naturally wonders whether a similar RKHS approach can be extended to sufficient dimension reduction such as functional sliced inverse regression (SIR), which was also traditionally based on functional PCA which assumes that the projection direction of interest is well-represented by the basis obtained from functional PCA. It is interesting to see whether the more general framework can lead to better performance in functional SIR.

Appendix: Proofs

Proof of Proposition 2. Let $\{\omega_j\}_{j=1}^{\infty}$ be the eigenfunctions of K corresponding to the eigenvalues $\alpha_1 \geq \alpha_2 \geq \cdots > 0$. Since $\beta(t,\cdot) \in H_K$, we can write $\beta(t,s) = \sum_j a_j(t)\omega_j(s)$ for some function a_j, with $\sum_{j=1}^{\infty} a_j^2(t)/\alpha_j < \infty$ (pointwise summable in t). For any $f = \sum_j f_j\omega_j \in H_K$, $BK^{-1/2}f = \sum_j (f_j/\sqrt{\alpha_j})a_j$. Using this representation, $BK^{-1/2}$ can be natually extended to L_2 by defining $BK^{-1/2}f = \sum_j (f_j/\sqrt{\alpha_j})a_j \in L_2$ for any $f \in L_2$. Using Cauchy-Schwartz inequality, this operator is obviously bounded on L_2 since the assumption that $\|\beta(t,\cdot)\|_{H_K} \in L_2$ implies $(\sum_j a_j^2/\alpha_j)^{1/2} \in L_2$. □

Proof of Theorem 1. In the proofs we use C to denote a generic positive constant. Using $\beta(t,\cdot) = K^{1/2} f(t,\cdot)$, from (4),

$$\hat{B}(X^*) = \frac{\sum_i (Y_i \otimes K^{1/2} X_i)}{n} (T_n + \lambda I)^{-1} K^{1/2} X^*,$$

where I is the identity operator, $T_n = K^{1/2} \Gamma_n K^{1/2}$ and $\Gamma_n = \sum_i (X_i \otimes X_i)/n$ is the empirical version of Γ. Using $Y_i = B(X_i) + \epsilon_i$, and noting that $T_n = \sum_i (K^{1/2} X_i \otimes K^{1/2} X_i)/n$, we have

$$\begin{aligned}
& \hat{B}(X^*) - B(X^*) \\
= \; & \frac{\sum_i (B(X_i) \otimes K^{1/2} X_i)}{n} (T_n + \lambda I)^{-1} K^{1/2} X^* + \frac{\sum_i (\epsilon_i \otimes K^{1/2} X_i)}{n} (T_n + \lambda I)^{-1} K^{1/2} X^* - B(X^*) \\
= \; & B K^{-1/2} \left(T_n (T_n + \lambda I)^{-1} - I \right) K^{1/2} X^* + \frac{\sum_i (\epsilon_i \otimes K^{1/2} X_i)}{n} (T_n + \lambda I)^{-1} K^{1/2} X^* \\
= \; & -\lambda B K^{-1/2} (T_n + \lambda I)^{-1} K^{1/2} X^* + \frac{\sum_i (\epsilon_i \otimes K^{1/2} X_i)}{n} (T_n + \lambda I)^{-1} K^{1/2} X^* \\
=: \; & A_1 + A_2.
\end{aligned}$$

We first deal with A_1. Note $A_1 = -\lambda B K^{-1/2} (T + \lambda I)^{-1} K^{1/2} X^* - \lambda B K^{-1/2} (T_n + \lambda I)^{-1} (T - T_n)(T + \lambda I)^{-1} K^{1/2} X^*$.

Using the expansion $K^{1/2} X^* = \sum_j \xi_j^* e_j$,

$$\begin{aligned}
& \lambda^2 E^* \| B K^{-1/2} (T + \lambda I)^{-1} K^{1/2} X^* \|^2 \\
= \; & \lambda^2 E^* \left[\sum_k \langle B K^{-1/2} \sum_j \frac{\xi_j^*}{s_j + \lambda} e_j, e_k \rangle^2 \right] \\
= \; & \lambda^2 \sum_{j,k} \frac{s_j}{(s_j + \lambda)^2} \langle B K^{-1/2} e_j, e_k \rangle \\
\leq \; & \frac{\lambda}{4} \sum_{j,k} \langle B K^{-1/2} e_j, e_k \rangle^2 \\
= \; & \frac{\lambda}{4} \| B K^{-1/2} \|_{HS}^2. \qquad (6)
\end{aligned}$$

Also, writing $\mathcal{A} = BK^{-1/2}(T_n+\lambda I)^{-1}(T-T_n)(T+\lambda I)^{-1}$ for simplicity of notation,

$$\begin{aligned}
& \lambda^2 E^* \|BK^{-1/2}(T_n+\lambda I)^{-1}(T-T_n)(T+\lambda I)^{-1}K^{1/2}X^*\|^2 \\
= \;& \lambda^2 E^* \langle \mathcal{A}K^{1/2}X^*, \mathcal{A}K^{1/2}X^* \rangle \\
= \;& \lambda^2 E^* \langle \mathcal{A}^{\mathrm{T}}\mathcal{A}K^{1/2}X^*, K^{1/2}X^* \rangle \\
= \;& \lambda^2 \mathrm{Trace}(\mathcal{A}^{\mathrm{T}}\mathcal{A}T) \\
= \;& \lambda^2 \|\mathcal{A}T^{1/2}\|_{HS}^2 \\
\leq \;& \lambda^2 \|BK^{-1/2}\|_{HS}^2 \|(T_n+\lambda I)^{-1}(T-T_n)(T+\lambda I)^{-1}T^{1/2}\|_{HS}^2 \\
\leq \;& \lambda^2 \|BK^{-1/2}\|_{HS}^2 \|(T_n+\lambda I)^{-1}\|_{op}^2 \|(T-T_n)(T+\lambda I)^{-1}T^{1/2}\|_{HS}^2 \\
= \;& O_p(\|(T-T_n)(T+\lambda I)^{-1}T^{1/2}\|_{HS}^2). \qquad (7)
\end{aligned}$$

We have

$$\begin{aligned}
& E\|(T-T_n)(T+\lambda I)^{-1}T^{1/2}\|_{HS}^2 \\
= \;& E\sum_{j,k}\langle (T-T_n)(T+\lambda I)^{-1}T^{1/2}e_j, e_k\rangle^2 \\
= \;& E\sum_{j,k}\langle (T-T_n)\frac{s_j^{1/2}}{(s_j+\lambda)}e_j, e_k\rangle^2. \qquad (8)
\end{aligned}$$

Direct calculation reveals that

$$\begin{aligned}
& E\langle (T-T_n)e_j, e_k\rangle^2 \\
= \;& E\langle s_j e_j - \frac{1}{n}\sum_i ((\sum_l \xi_{il}e_l)\otimes(\sum_m \xi_{im}e_m))e_j, e_k\rangle^2 \\
= \;& E\langle s_j e_j - \sum_i \frac{\sum_l \xi_{il}\xi_{ij}e_l}{n}, e_k\rangle^2 \\
= \;& E(s_j I\{j=k\} - \frac{\sum_i \xi_{ij}\xi_{ik}}{n})^2 \\
\leq \;& E(\frac{\sum_i \xi_{ij}\xi_{ik}}{n})^2,
\end{aligned}$$

where the last step used the fact that $E[\xi_{ij}\xi_{ik}] = s_j I\{j=k\}$. Using assumption (A2),

we have $E\langle (T-T_n)e_j, e_k\rangle^2 \le Cs_j s_k/n$, which combined with (8) implies

$$E\|(T-T_n)(T+\lambda I)^{-1}T^{1/2}\|_{HS}^2 \le \frac{C}{n}\sum_{j,k}\frac{s_j^2 s_k}{(s_j+\lambda)^2} = O\left(\frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2}\right). \quad (9)$$

(6),(7) and (9) together yield $E^*\|A_1\|^2 = O_p(\lambda + \frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2})$.

Now, write $A_2 = \frac{\sum_i(\epsilon_i \otimes K^{1/2}X_i)}{n}(T+\lambda I)^{-1}K^{1/2}X^* - \frac{\sum_i(\epsilon_i\otimes K^{1/2}X_i)}{n}(T+\lambda I)^{-1}(T - T_n)(T_n+\lambda I)^{-1}K^{1/2}X^*$. We have

$$\begin{aligned} & E^*\|\frac{\sum_i(\epsilon_i\otimes K^{1/2}X_i)}{n}(T+\lambda I)^{-1}K^{1/2}X^*\|^2 \\ = \;& E^*\|\frac{1}{n}\sum_i \epsilon_i\langle K^{1/2}X_i, \sum_j \frac{\xi_j^*}{s_j+\lambda}e_j\rangle\|^2, \end{aligned}$$

and thus

$$\begin{aligned} & E\|\frac{\sum_i(\epsilon_i\otimes K^{1/2}X_i)}{n}(T+\lambda I)^{-1}K^{1/2}X^*\|^2 \\ = \;& \frac{\sigma_\epsilon^2}{n}E\left[\langle K^{1/2}X_1, \sum_j \frac{\xi_j^*}{s_j+\lambda}e_j\rangle^2\right] \\ = \;& \frac{\sigma_\epsilon^2}{n}E\sum_j \frac{\xi_{1j}^2\xi_j^{*2}}{(s_j+\lambda)^2} \\ = \;& \frac{\sigma_\epsilon^2}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2}. \end{aligned}$$

where $\sigma_\epsilon^2 = E\|\epsilon\|^2$. Furthermore, denoting $\mathcal{C} = (T+\lambda I)^{-1}(T-T_n)(T_n+\lambda I)^{-1}$,

$$
\begin{aligned}
& E\left[\|\frac{\sum_i(\epsilon_i \otimes K^{1/2}X_i)}{n}(T_n+\lambda I)^{-1}(T-T_n)(T+\lambda I)^{-1}K^{1/2}X^*\|^2|X_1,\ldots,X_n\right] \\
= & \frac{\sigma_\epsilon^2}{n^2}E\left[\sum_i\langle K^{1/2}X_i,(T_n+\lambda I)^{-1}(T-T_n)(T+\lambda I)^{-1}K^{1/2}X^*\rangle^2|X_1,\ldots,X_n\right] \\
= & \frac{\sigma_\epsilon^2}{n^2}E\left[\sum_i\langle \mathcal{C}K^{1/2}X_i,K^{1/2}X^*\rangle^2|X_1,\ldots,X_n\right] \\
= & \frac{\sigma_\epsilon^2}{n^2}\left[\sum_i\langle \mathcal{C}^{\mathrm{T}}T\mathcal{C}K^{1/2}X_i,K^{1/2}X_i\rangle\right] \\
= & \frac{\sigma_\epsilon^2}{n}\mathrm{Trace}(\mathcal{C}^{\mathrm{T}}T\mathcal{C}T_n) \\
= & \frac{\sigma_\epsilon^2}{n}\mathrm{Trace}(T_n^{1/2}\mathcal{C}^{\mathrm{T}}T^{1/2}T^{1/2}\mathcal{C}T_n^{1/2}) \\
= & \frac{\sigma_\epsilon^2}{n}\|T_n^{1/2}\mathcal{C}^{\mathrm{T}}T^{1/2}\|_{HS}^2 \\
= & \frac{\sigma_\epsilon^2}{n}\|T_n^{1/2}(T_n+\lambda I)^{-1}(T-T_n)(T+\lambda I)^{-1}T^{1/2}\|_{HS}^2 \\
\leq & \frac{\sigma_\epsilon^2}{n}\|T_n^{1/2}(T_n+\lambda I)^{-1}\|_{op}^2\|(T-T_n)(T+\lambda I)^{-1}T^{1/2}\|_{HS}^2 \\
= & O_p(\frac{1}{n\lambda})\cdot O_p(\frac{1}{n}\sum_j\frac{s_j^2}{(s_j+\lambda)^2}) \\
= & o_p(\frac{1}{n}\sum_j\frac{s_j^2}{(s_j+\lambda)^2}),
\end{aligned}
$$

where we used (9) and that $n\lambda \to \infty$. Thus we have $E^*\|A_2\|^2 = O_p(\frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2})$. The theorem is proved by combining the bounds for $E^*\|A_1\|^2$ and $E^*\|A_2\|^2$. □

Proof of Theorem 2. Our model is $Y(t) = \int \beta(t,s)X(s)ds + \epsilon(t)$. Consider the special case $\beta(t,s) = e_1(t)\otimes\beta(s)$ and $\epsilon(t) = e_1(t)\chi$, where $\beta(s) \in H_K$, $\|\beta\|_{H_K} \leq 1$, and $\chi \sim N(0,\sigma^2)$. Then by taking inner products with e_j on both sides of $Y(t) = \int \beta(t,s)X(s)ds + \epsilon(t)$, the model becomes $Y^{(1)} = \int \beta(s)X(s)ds + \chi$, $Y^{(2)} = Y^{(3)} = \cdots = 0$ where $Y^{(j)} = \langle Y, e_j\rangle$. Since $\|\int\beta(\cdot,s)X(s)ds\| = |\int\beta(s)X(s)ds|$, the lower bound for the scalar response model provides a lower bound for the functional response

model. Thus we can just consider the model with scalar response:

$$Y_i = \int \beta(s) X_i(s) ds + \chi_i,$$

with $\|\beta\|_{H_K} \leq 1$. We need a modification of the proof of Theorem 1 in Cai and Yuan (2012) due to the more general assumption on the eigenvalues of T. Let $\eta_j = \sqrt{c\lambda_0/(J_0 s_j)}$ for some $0 < c \leq 1$ to be determined later. We apply Theorem 2.5 of Tsybakov (2009) using the following collection of 2^{J_0} functions

$$f_\theta = \sum_{k=1}^{J_0} \theta_k \eta_k K^{1/2} e_k,$$

where $\theta = (\theta_1, \ldots, \theta_{J_0}) \in \{0,1\}^{J_0}$.

First, using that $\|K^{1/2} e_j, K^{1/2} e_k\|_{H_K} = \langle e_j, e_k \rangle = 1\{j = k\}$,

$$\|f_\theta\|_{H_K}^2 = \sum_{k=1}^{J_0} \theta_k^2 \eta_k^2 \leq \sum_{k=1}^{J_0} \eta_k^2 = \frac{c\lambda_0}{J_0} \sum_{k=1}^{J_0} \frac{1}{s_k} \leq \frac{c\lambda_0}{J_0} \frac{J_0}{s_{J_0}} \leq c \leq 1,$$

since $s_{J_0} \geq \lambda_0$ by $s_{J_0} = \phi(J_0)$ and the definition $J_0 = \lfloor \phi^{-1}(\lambda_0) \rfloor$.

By the Varshamov-Gilbert bound (Lemma 2.9 in Tsybakov (2009)), there is a subset $\Theta = \{\theta^0, \ldots, \theta^N\} \subset \{0,1\}^{J_0}$ such that $\theta^0 = (0, \ldots, 0)$, $N \geq 2^{J_0/8}$ and $\sum_{k=1}^{J_0} (\theta_k - \theta'_k)^2 \geq J_0/8$ whenever $\theta \neq \theta' \in \Theta$.

We have

$$\|\Gamma^{1/2}(f_\theta - f_{\theta'})\|^2 = \sum_{k=1}^{J_0} \eta_k^2 (\theta_k - \theta'_k)^2 s_k \geq \frac{c\lambda_0}{J_0} \frac{J_0}{8} = c\lambda_0/8,$$

verifying condition (i) in Theorem 2.5 of Tsybakov (2009). Furthermore, the Kullback-Leibler distance between P_θ and $P_{\theta'}$ (P_θ is the joint distribution of training data when $\beta = f_\theta$) can be found to be

$$K(P_\theta | P_{\theta'}) = \frac{n}{2\sigma^2} \sum_{k=1}^{J_0} \eta_k^2 (\theta_k - \theta'_k)^2 s_k \leq \frac{nc\lambda_0}{2\sigma^2},$$

and thus

$$\frac{1}{N}\sum_{j=1}^{N} K(P_\theta|P_{\theta'}) \leq \frac{nc\lambda_0}{2\sigma^2} = \frac{c\phi^{-1}(\lambda_0)}{2\sigma^2} \leq \frac{c}{2\sigma^2}(J_0+1) \leq \alpha \log N,$$

for some $0 < \alpha < 1/8$ if c is chosen small enough, verifying condition (ii) in Theorem 2.5 of Tsybakov (2009). The lower bound is proved by applying Theorem 2.5 of Tsybakov (2009). □

References

Aguilera, A., Ocana, F., and Valderrama, M. "Estimation of functional regression models for functional responses by wavelet approximation." *International Workshop on Functional and Operatorial Statistics* (2008).

Ait-Saidi, A., Ferraty, F., Kassa, R., and Vieu, P. "Cross-validated estimations in the single-functional index model." *Statistics*, 42(6):475–494 (2008).

Antoch, J., Prchal, L., De Rosa, M. R., and Sarda, P. "Functional linear regression with functional response: Application to prediction of electricity consumption." *International Workshop on Functional and Operatorial Statistics* (2008).

Cai, T. and Hall, P. "Prediction in functional linear regression." *Annals of Statistics*, 34(5):2159–2179 (2006).

Cai, T. and Yuan, M. "Minimax and adaptive prediction for functional linear regression." *Journal of the American Statistical Association*, 107(499):1201–1216 (2012).

Cardot, H., Ferraty, F., and Sarda, P. "Spline estimators for the functional linear model." *Statistica Sinica*, 13(3):571–591 (2003).

Cardot, H., Mas, A., and Sarda, P. "CLT in functional linear regression models." *Probability Theory and Related Fields*, 138(3):325–361 (2007).

Crambes, C., Kneip, A., and Sarda, P. "Smoothing splines estimators for functional linear regression." *Annals of Statistics*, 37(1):35–72 (2009).

Crambes, C. and Mas, A. "Asymptotics of prediction in functional linear regression with functional outputs." *Bernoulli*, to appear (2012).

Ferraty, F., González-Manteiga, W., Mart´nez-Calvo, A., and Vieu, P. "Presmoothing in functional linear regression." *Statistica Sinica*, 22:69–94 (2011).

Ferraty, F. and Vieu, P. "The functional nonparametric model and application to spectrometric data." *Computational Statistics*, 17(4):545–564 (2002).

Hall, P. and Horowitz, J. L. "Methodology and convergence rates for functional linear regression." *Annals of Statistics*, 35(1):70–91 (2007).

Lian, H. "Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces." *Canadian Journal of Statistics-Revue Canadienne De Statistique*, 35(4):597–606 (2007).

—. "Convergence of functional k-nearest neighbor regression estimate with functional responses." *Electronic Journal of Statistics*, 5:31–40 (2011).

Preda, C. "Regression models for functional data by reproducing kernel Hilbert spaces methods." *Journal of Statistical Planning and Inference*, 137(3):829–840 (2007).

Ramsay, J. O. and Silverman, B. W. *Functional data analysis*. Springer series in statistics. New York: Springer, 2nd edition (2005).

Tsybakov, A. *Introduction to Nonparametric Estimation*. New York: Springer (2009).

Wahba, G. *Spline models for observational data*. Philadelphia, PA: Society for Industrial and Applied Mathematics (1990).

Wong, H., Zhang, R. Q., Ip, W. C., and Li, G. Y. "Functional-coefficient partially linear regression model." *Journal of Multivariate Analysis*, 99(2):278–305 (2008).

Yao, F., Mueller, H. G., and Wang, J. L. "Functional linear regression analysis for longitudinal data." *Annals of Statistics*, 33(6):2873–2903 (2005).

Yuan, M. and Cai, T. T. "A reproducing kernel Hilbert space approach to functional linear regression." *Annals of Statistics*, 38(6):3412–3444 (2010).